CALIFORNIA
ENVIRONMENTAL
HEALTH TRACKING
PROGRAM

Spatial Analytic Modeling of the Health Effects of Traffic Emissions

- Eric Roberts, MD PhD
- Paul English, PhD MPH
- Craig Wolf, MS-Eng
- Makinde Falade, MS-GIS
- Svetlana Smorodinsky, MPH

What are the health-hazard associations in my data?

- Health outcome: ER visits due to asthma among Alameda County, California residents enrolled in Medicaid and Kaiser Permanente; geocoded by residence address
- Hazard metric: Annualized average vehicle counts within a 300-meter radius of subject's residence

The approach I learned in grad school:

$$f(ER) = b_0 + b_{\exp} x_{\exp} + b_{\text{cov}} x_{\text{cov}} + \sum e_i$$

- *f(ER)* is a function of the ER visit rate, such as the log or logit
- x_{exp} is my exposure metric(s)
- x_{cov} are my important covariates such as race or social class; these may not be of direct interest but their inclusion can profoundly influence the other parameters I calculate

The approach I learned in grad school:

$$f(ER) = b_0 + b_{\exp} x_{\exp} + b_{\text{cov}} x_{\text{cov}} + \sum e_i$$

- β_{exp} are the parameters that I'm really interested in – they represent the associations between traffic and ER visits
- ε_i are my residuals – in order for my parameter estimates to be valid, these must be evenly distributed around zero and their mean must be constant throughout Alameda County

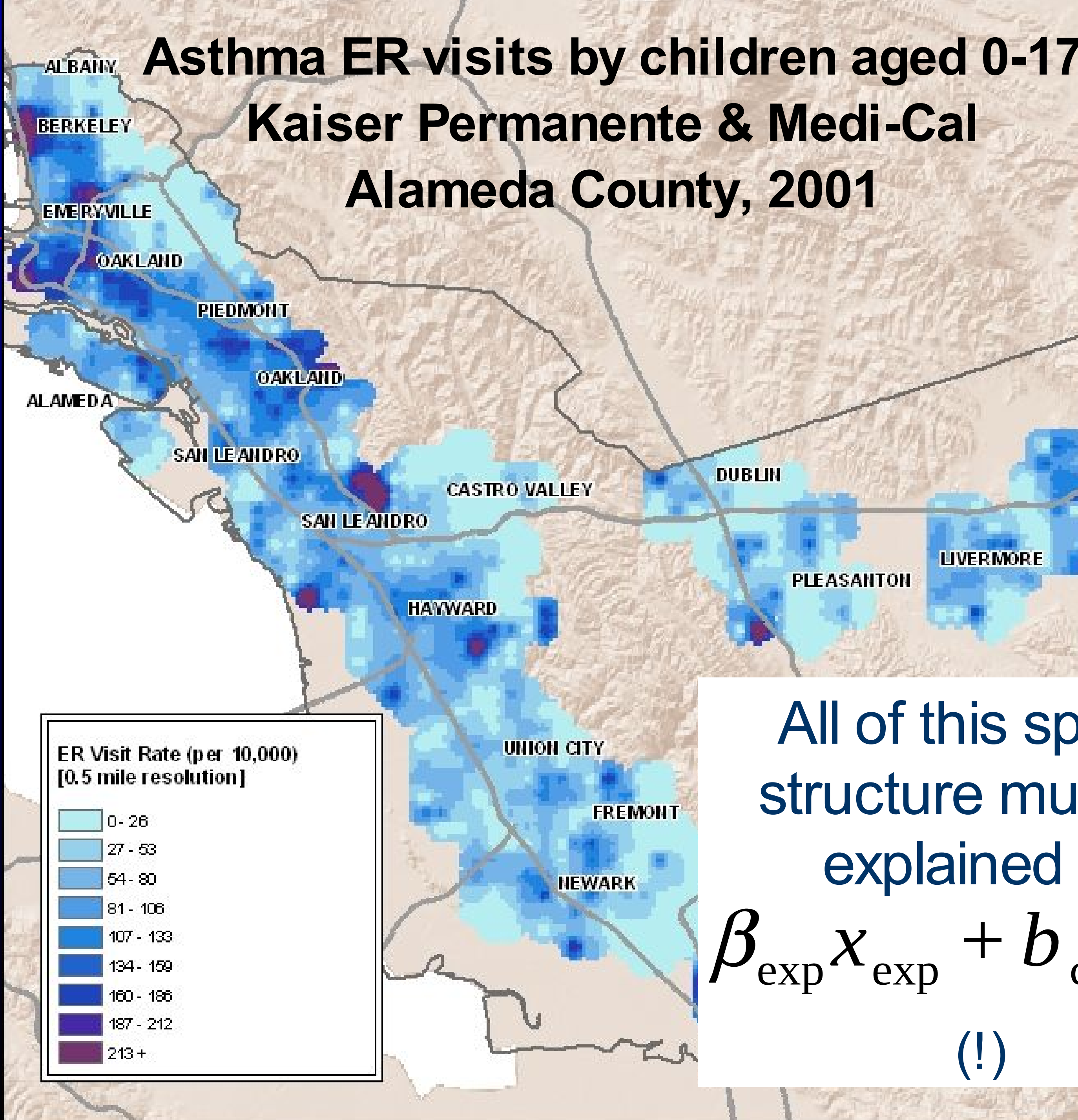
Traditional approach

- Advantages:
  1. Understood by all professional audiences
  2. Common understanding about how regression results behave under different circumstances
  3. Many accessible software packages to choose from

Traditional approach

- Biggest disadvantage:
  - My assumption about the constant mean of the residuals (i.e. that all spatial structure is accounted for by x_{exp} and x_{cov}) is extremely tenuous)

Asthma ER visits by children aged 0-17
Kaiser Permanente & Medi-Cal
Alameda County, 2001



All of this spatial structure must be explained by

$$\beta_{\exp} x_{\exp} + b_{\text{cov}} x_{\text{cov}}$$

(!)

One solution

$$f(ER) = b_0 + b_{\exp} x_{\exp} + b_{\mathrm{cov}} x_{\mathrm{cov}} + Sp(x, y)$$

- Allows for the spatial structure of your residuals in your regression model
- Any residual geographic variation not accounted for by your parameters is represented in the residual structure (“residual variation”)

What can we use for *Sp(x,y)*?

- Several options, although more have been developed for aggregated (regional) rather than point data
- Could be a description of how the covariance between neighboring points decreases with distance (Kriging)
- Another option is *locally weighted estimation* (loess) – the regression (usually linear or quadratic) of the outcome *f(ER)* as a function of the coordinates across the space

Implications of including non-parametric terms

- *lo(x,y)* will take on a different form depending on what you leave out of your model; in

$$f(ER) = \beta_0 + lo(x, y)$$

  lo(x,y) just depicts the spatial variation of *f(ER)*

- If *lo(x,y)* adapts to fit whatever is left out of the model, will β_{exp} change depending on whether we include our covariates anymore?

$$f(ER) = \beta_0 + \beta_{\text{exp}} x_{\text{exp}} + \beta_{\text{cov}} x_{\text{cov}} + lo(x, y)$$

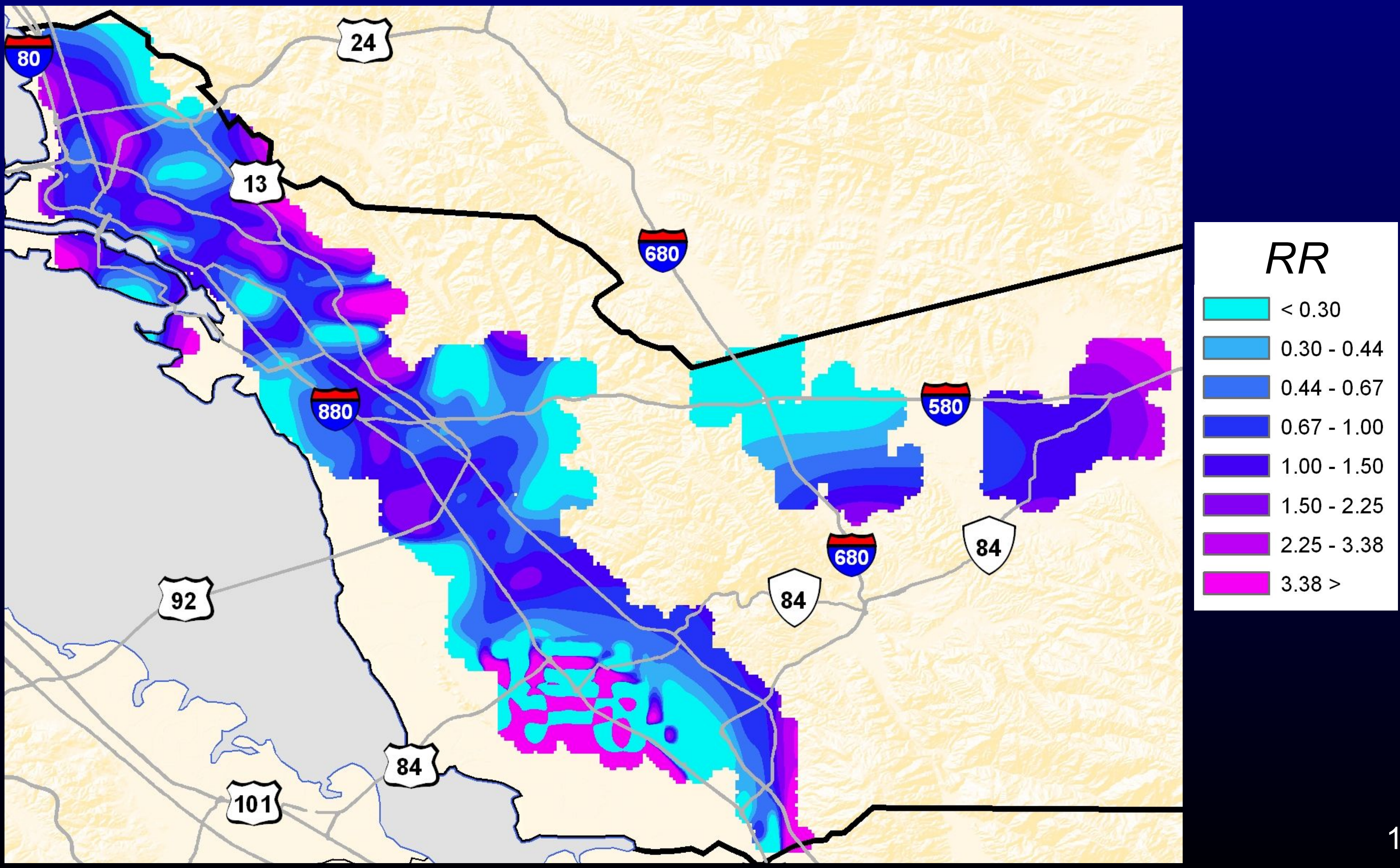
$$f(ER) = \beta_0 + \beta_{\text{exp}} x_{\text{exp}} + lo(x, y)$$

Enough talk already,
let’s give it a shot!

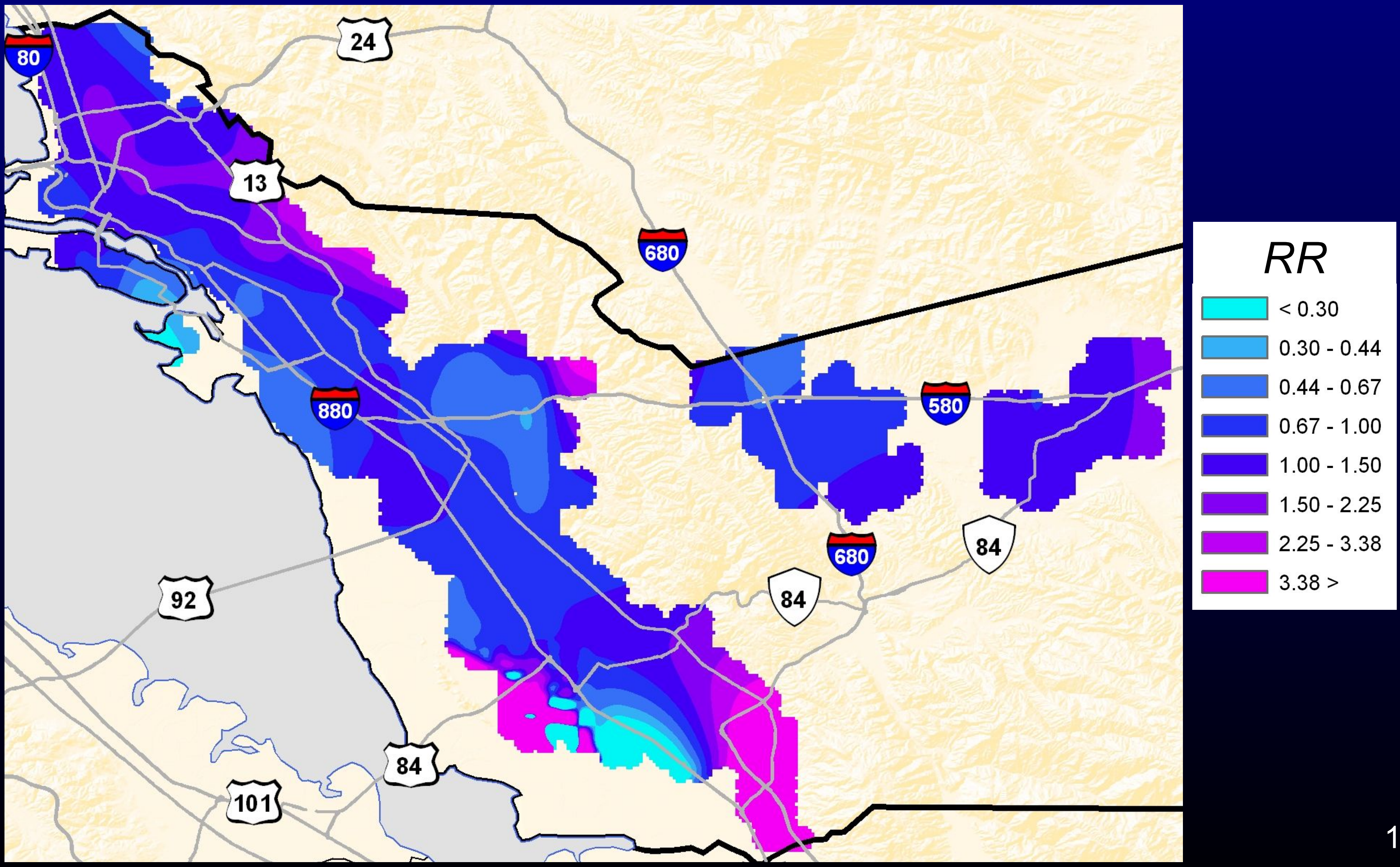
Estimating *lo(x,y)*

- At any given point, the nearest neighbors are included in the locally weighted estimation
- The fraction of all the data included as nearest neighbors is the *span*, which must be specified
- A small span gives a “bumpy” loess function, whereas a larger one will be smoother
- Still working out the bugs…

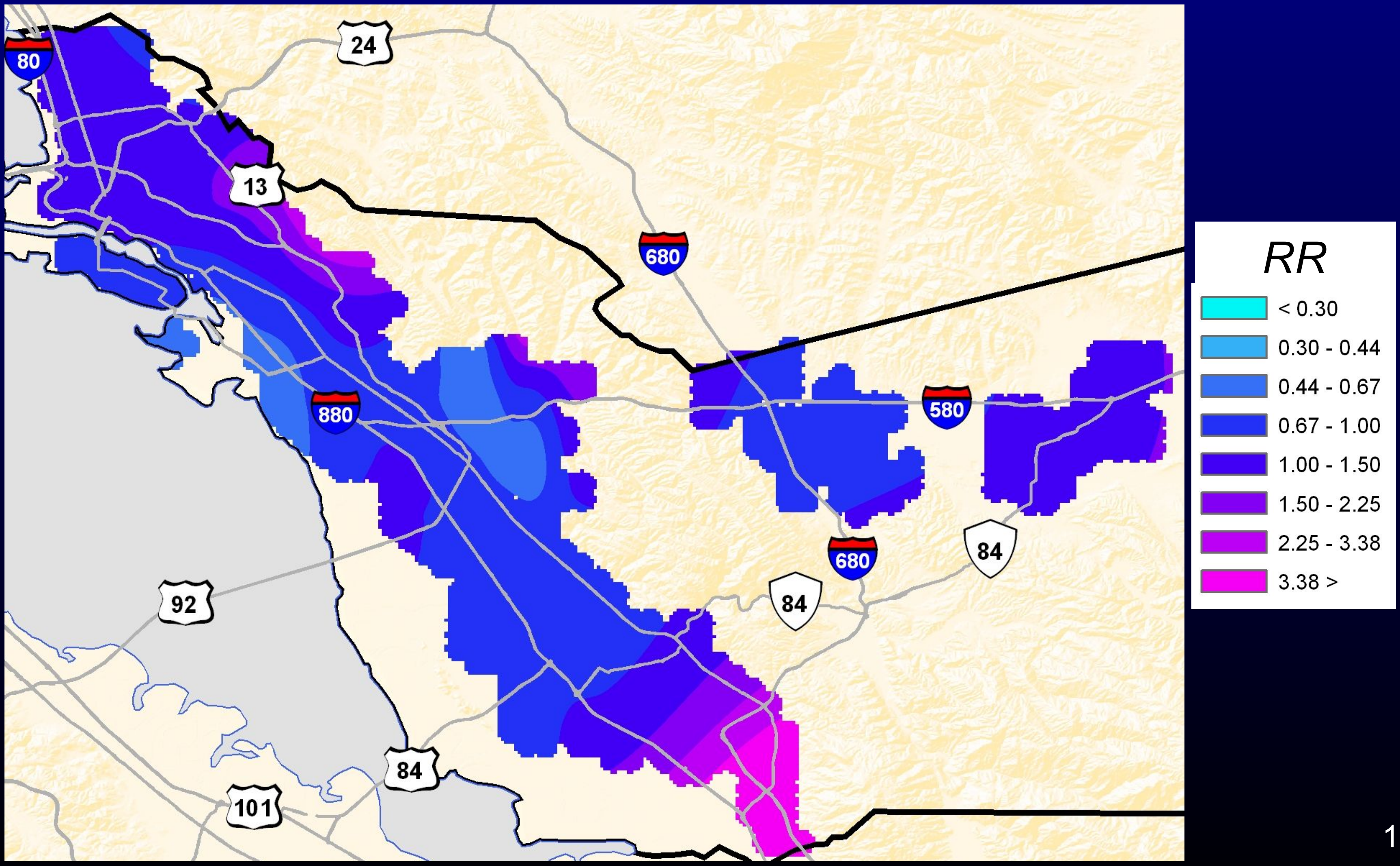
Relative risk for ER visits, children 0-4 span=0.04



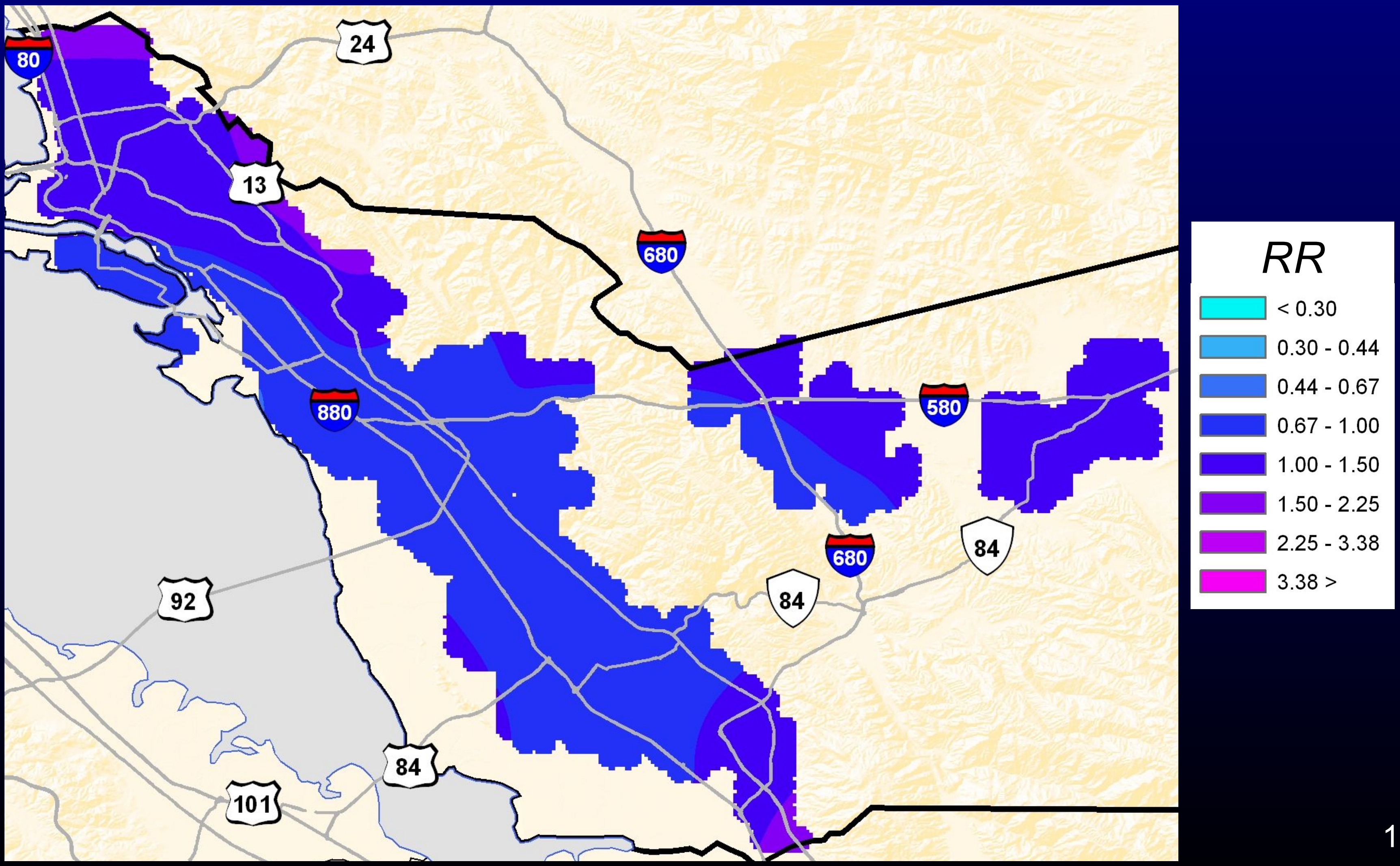
Relative risk for ER visits, children 0-4 span=0.15



Relative risk for ER visits, children 0-4 span=0.30

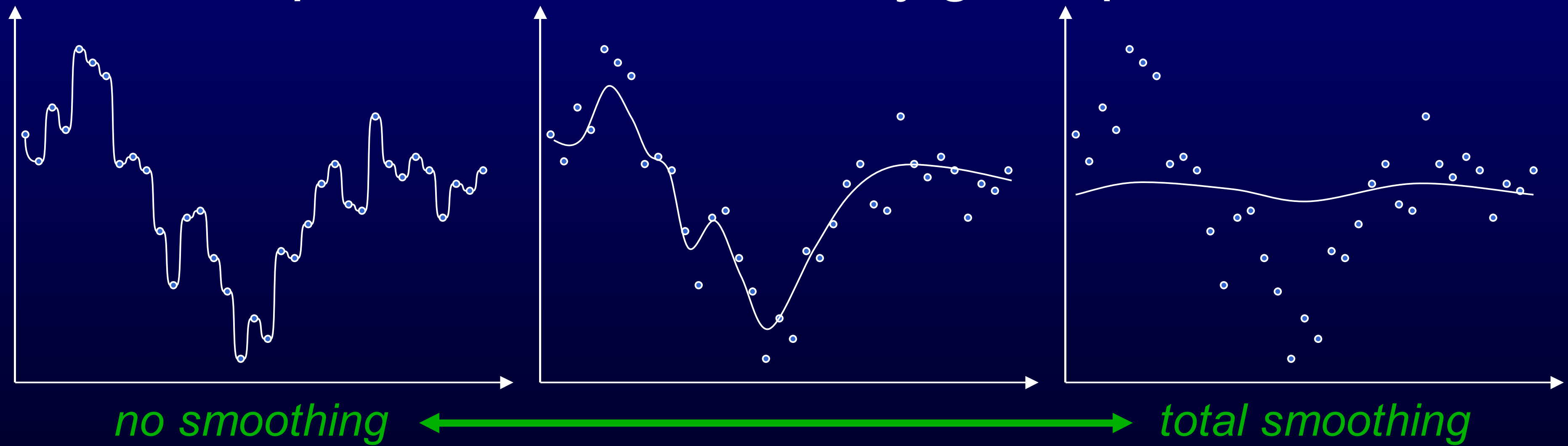


Relative risk for ER visits, children 0-4 span=0.60



How much smoothing is the "right amount?"

- This is really a question of how well the curve can predict the value of any given point



- Predictive power is maximized when Akaike's Information Criterion (AIC) is minimized
- Is this the best approach for all environmental epidemiology questions? All community stakeholder communications?

Finding the minimum AIC can be a chore!

- Large samples and/or large span values require heavy computing resources
- A random subset of the total sample may have a different span value associated with the AIC minimum

Optimal span sizes for asthma ER visit data

Age group	Available sample	Sample used	Span with AIC_{min}
0-4 years	44,526	(all)	0.078
5-17 years	78,433	25,000	1.000
18-44 years	197,856	25,000	0.077
45-64 years	111,235	25,000	0.141
65+ years	64,820	25,000	1.000

Does this change anything?

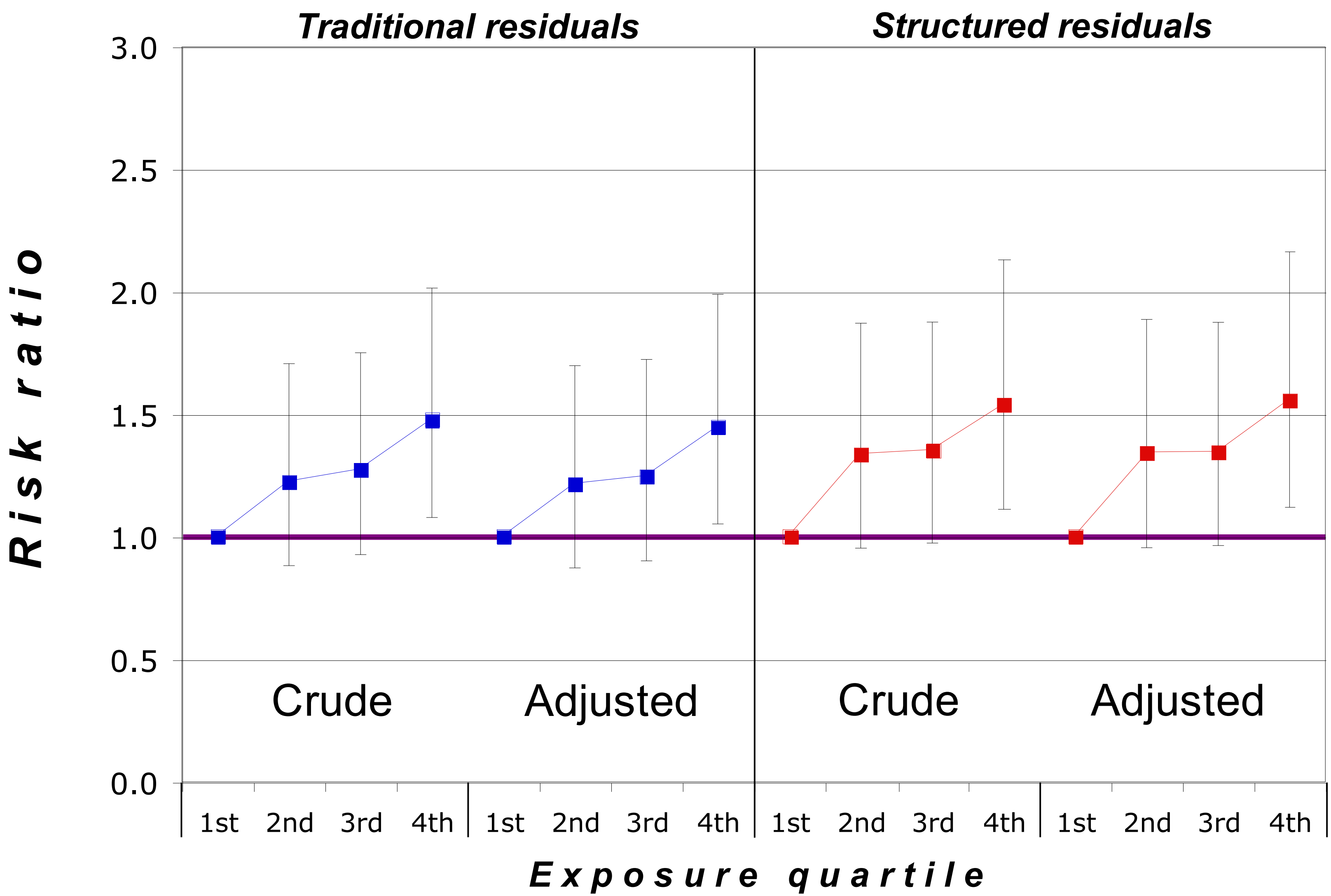
- Compared RR for ER visits among quartiles of exposure to our traffic metric (reference=1st quartile)

- Models:
  - Traffic only
  - Traffic + median family income of census tract + Medicaid status
  - Traffic ***with loess smooth term***
  - Traffic + median family income of census tract + Medicaid status ***with loess smooth term***

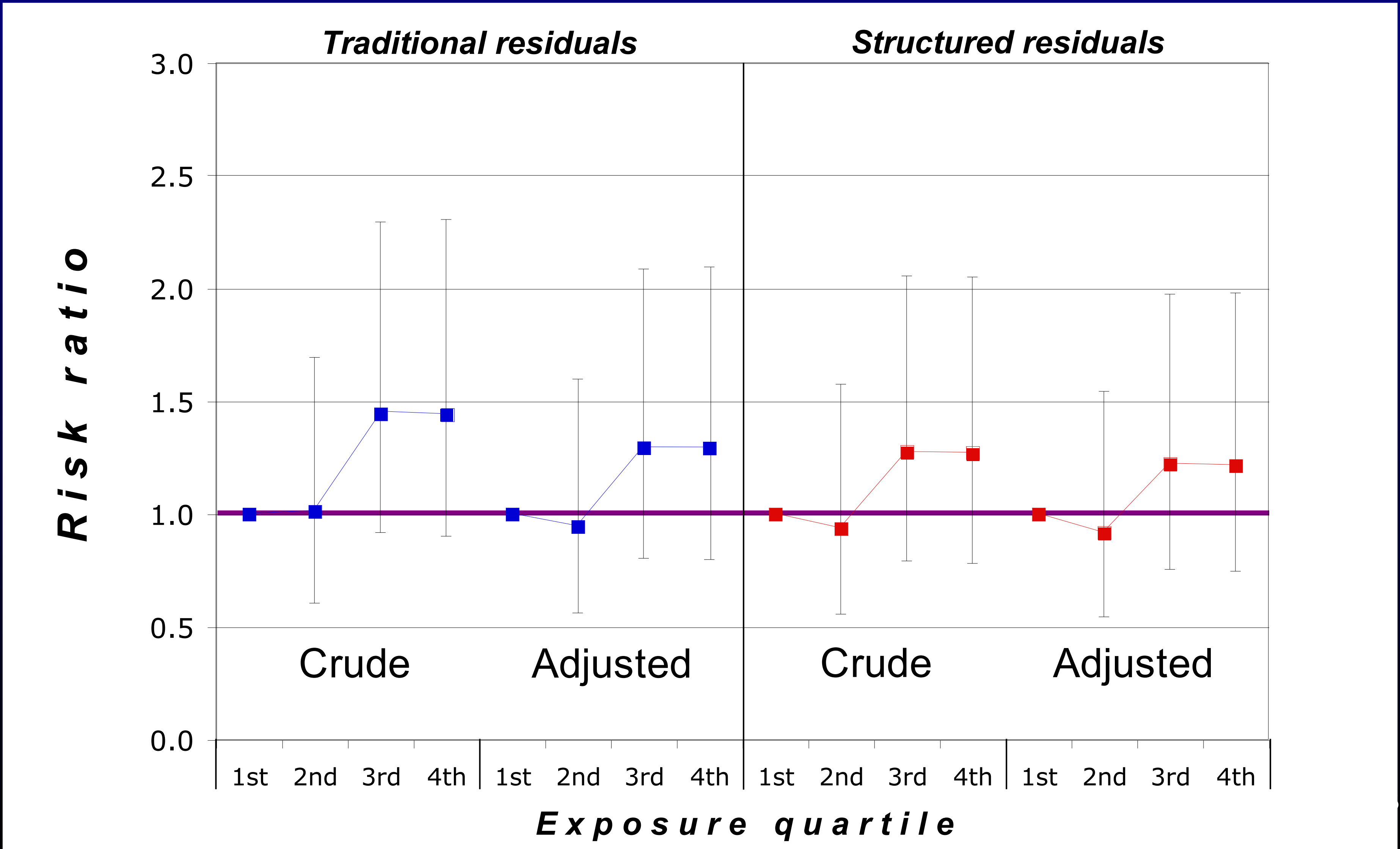
Results

- When smooth term included, the presence or absence of covariates in the model becomes mostly irrelevant *(for these data!)*
- Inclusion of the smooth term may make the pollution parameters…
  - More like the crude model
  - More like the adjusted model
  - Similar to neither model

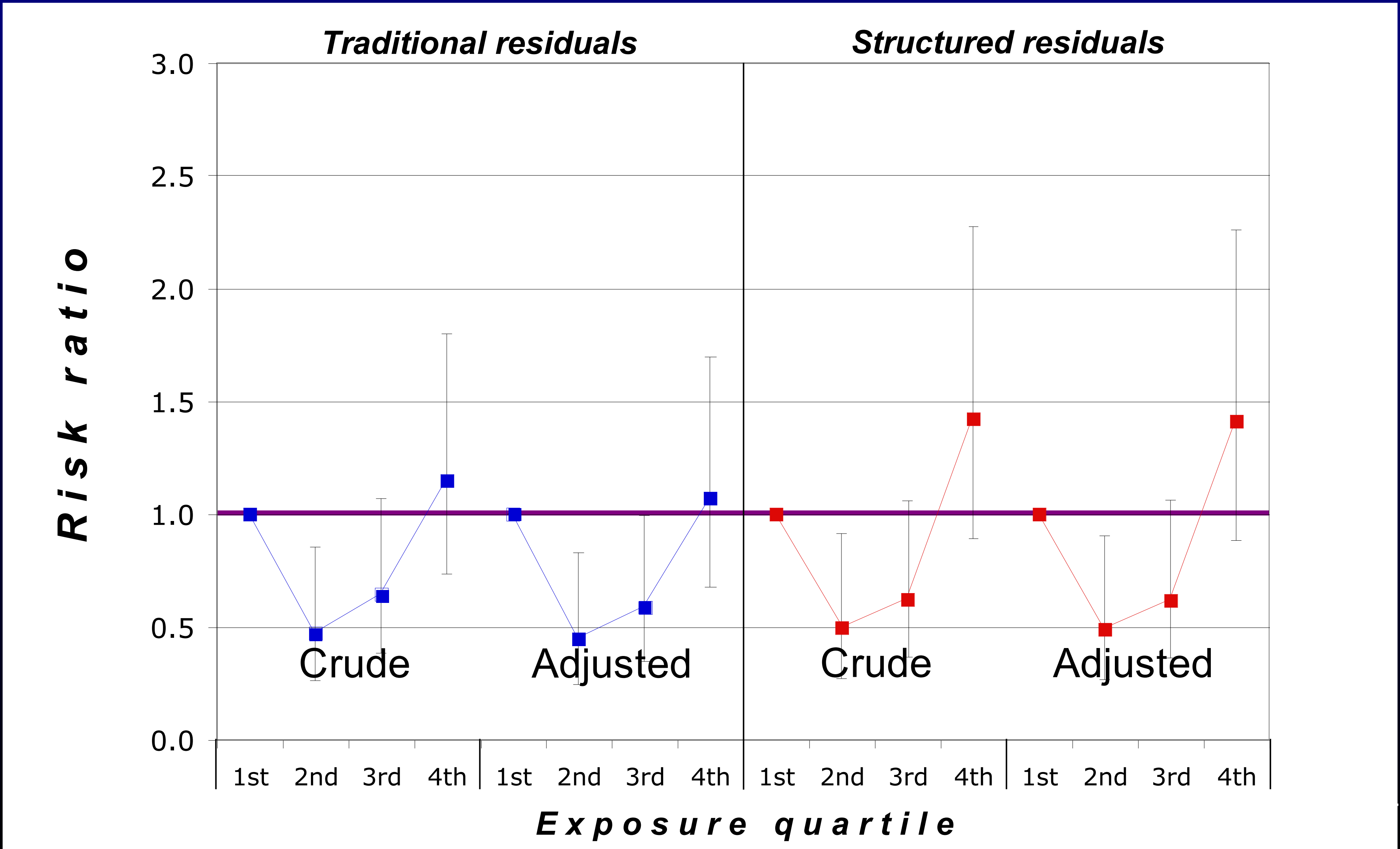
ER visits: Ages 0-4



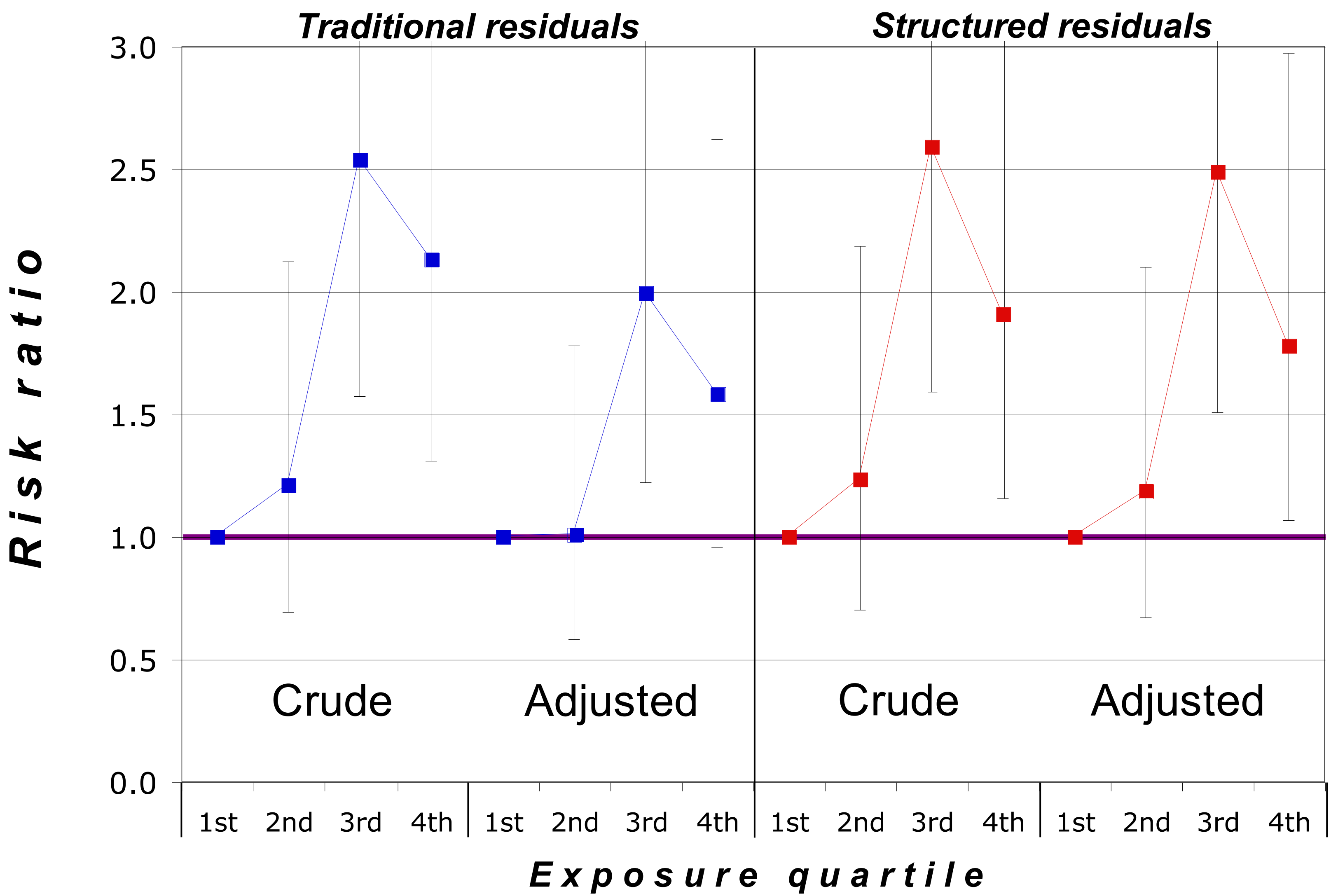
ER visits: Ages 5-17



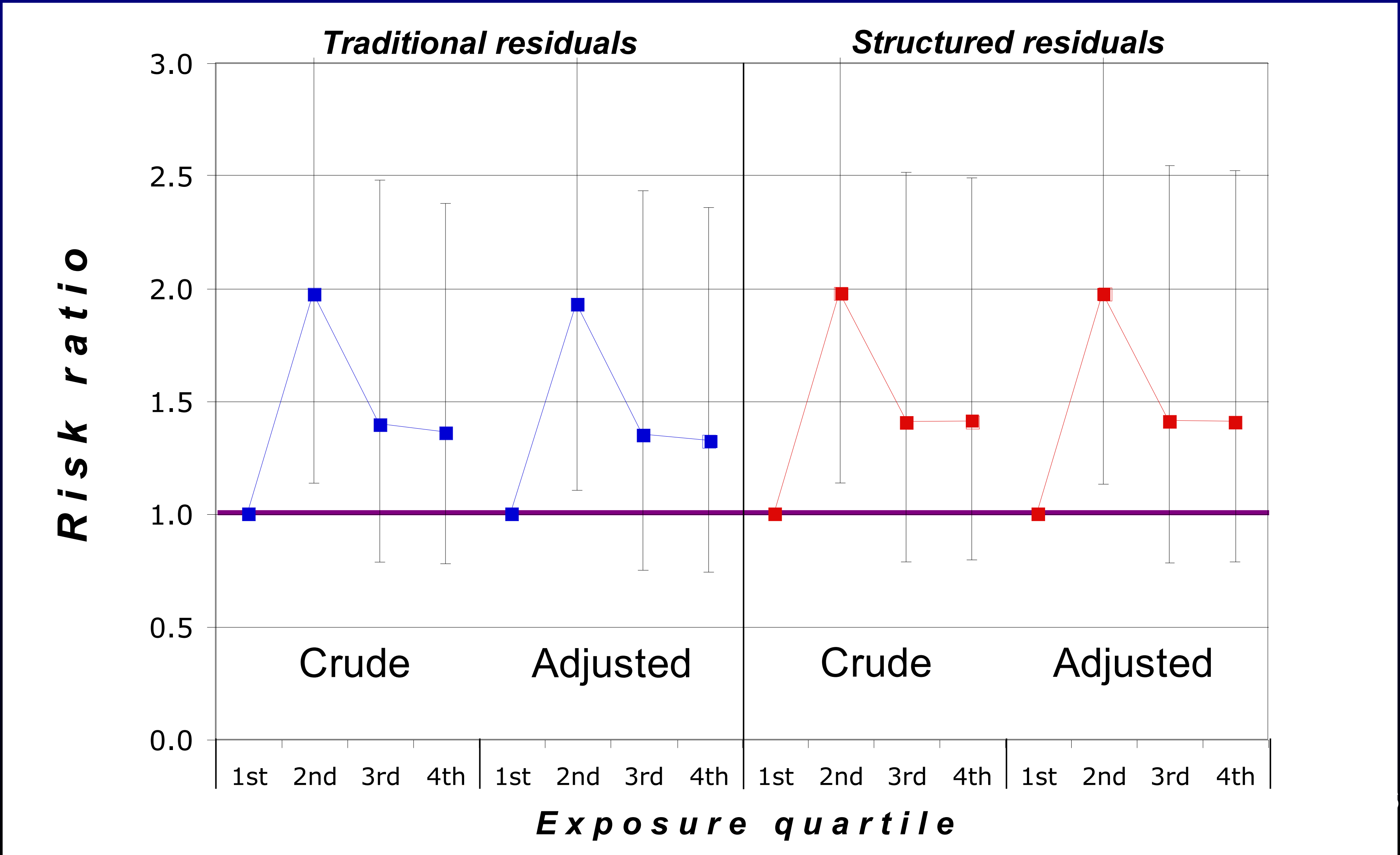
ER visits: Ages 18-44



ER visits: Ages 45-64



ER visits: Ages 65+



Lessons learned

- Working with spatially structured residuals is not too difficult
- Big step is selection of span value – need to explore appropriateness of AIC_{min} criterion for Tracking purposes
- We need much more experience to understand:
  - The relationship between sample size and the span value for AIC_{min}
  - How and when to graphically represent represent smoothed residuals
  - Under what circumstances it's ok to leave covariates out of our regression models

Lessons learned

- Often in Tracking we lack satisfactory covariate measures – if these can become irrelevant, this could be a big advantage
- Ability to visualize the residuals as maps may be a great tool on its own
  - Potentially objective criterion to determine appropriate resolution for maps
  - Ability to control for covariates to alter and view the resulting residual structure

Thank you!

The CDC Environmental Public Health Tracking Program

Lance A. Waller, PhD*

**For invaluable comments and advice; any errors and misconceptions are my fault, not his!*